FACTA UNIVERSITATIS (NIŠ)
Series: Electronics and Energetics vol.13, No.1, April 2000, 27–38

# INHABITED TV: MULTIMEDIA BROADCASTING FROM LARGE SCALE COLLABORATIVE VIRTUAL WORLD

**Milena Radenković, Chris Greenhalgn and Steve Benford**

**Abstract.** Inhabited TV is a new and exciting medium for entertainment and social communication which combines CVEs with broadcast TV [5]. This paper explores motivations for and the design of scaleable audio service for CVEs used in Inhabited TV. We proposed a flexible and scaleable distributed architecture for real-time voice mixing which is specifically tailored to support very dynamic sessions, rapidly varying requirements and very high levels of participation. We also consider use of crowd audio synthesis and environment sonification which can minimise network traffic in the audio medium, and thus further increasing its scalability.

**Key words:** Inhabited TV, collaborative virtual environments (CVE), scalable network architectures.

## 1. Introduction

This paper introduces the idea of Inhabited TV as a new medium for entertainment and social communication which combines collaborative virtual environments (CVEs) with broadcast TV. The fundamental idea of this medium is that on- line audience can participate in TV programs within shared virtual worlds. This extends traditional broadcast TV and more recent interactive TV by enabling *social interaction* among on-line audiences in the 3D shared virtual environments so that the audience becomes directly involved in the content of the show.

Manuscript received February 2, 2000. A version of this paper was presented at the fourth IEEE Conference on Telecommunications in Modern Satellite, Cables and Broadcasting Services, TELSIKS'99, October 1999, Niš, Sebia..

The authors are with School of Computer Science and IT University of Nottingham, Jubilee Campus, Willaton Road Nittigham, NG8 1BB England, e-mails are [mvr; cmg; sdb]@cs.nott.ac.uk.

Before we describe some experiments and issues raised, we introduce the idea of layered participation as a mechanism for describing Inhabited Television and defining associated terminology. Each layer provides different possibilities for navigation, interaction, mutual awareness and communication between participants, and is supported by a distinct combination of interface and transmission technologies. The layers are shown in the Fig. 1.

- The innermost layer describes the professional performers in the TV show. They have are embodied within the virtual worlds and have the fullest involvement in the show. In turn, this requires the support of the most powerful equipment such as immersive peripherals, high performance workstations and high-speed networks.
- The next layer describes online inhabitants who are also embodied within the virtual worlds. They are able to communicate with one another. Inhabitants use commonly navigate the virtual world, interact with its content and available equipment such as PCs, set-top boxes and public networks.
- The outermost layer describes viewers, who experience the show via broadcast or interactive TV and have most limited possibilities for the involvement and interaction
- The final layer is the production team, including camera operators, directors and technical support who works behind the scenes.

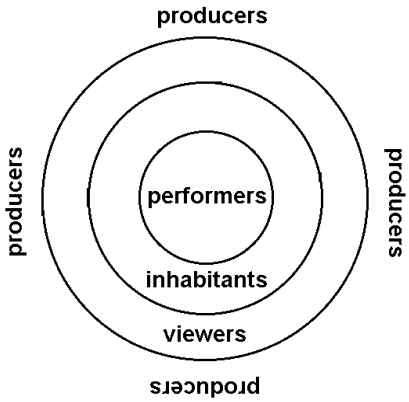


*Fig. 1. Layered participation in Inhabited TV.*

Experiments with Inhabited TV that we carried out in association with BT over the last few years are briefly described below. Each of the experiments conformed to our model of layered participation.

- *The Mirror* involved public access to a series of six public online virtual worlds on the Internet. The experiment was done in association with the BBC and Sony and run alongside UK BBC television series "The Net". The broadcast in this case was based on edited recorded footage.
- *Heaven & Hell-Live* was a live 1-hour television broadcast on the UK's Channel 4 inside a public virtual world. It was done in association with Sony and Illuminations TV). The overall participation structure of this show is given in the Fig. 2.
- Our most recent experiment was *Out of This World* (*OOTW*), a live-game show staged at multi-user virtual environment which was "broadcast" live onto a large screen in a theatre space. The participation structure of this experiment is shown in Fig. 3. The inhabitants were divided into two teams, aliens and robots, who had to race across a doomed space station in order to reach the one remaining escape craft. On their way they had to compete in a series of interactive games and collaborative tasks in order to score points. The following figures show various scenes from the show. Fig. 4 shows the Alien team captain and team members. Fig. 5 shows the robot team captain on his platform in the falling fish game.

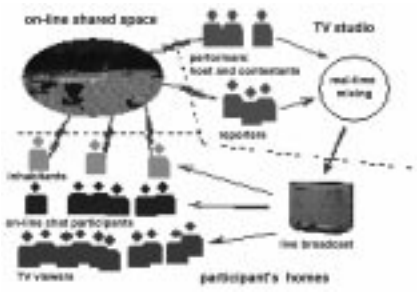


*Fig. 2. Layers of participation in Heaven & Hell-Live*

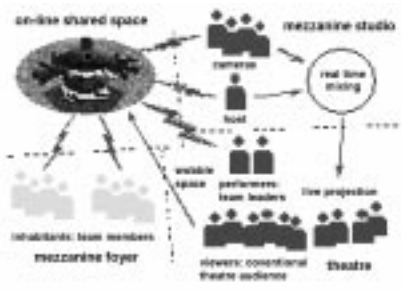


*Fig. 3. Participation structure of OOTW*

The two teams each consisted of four inhabitants, members of the public who had been selected from the theatre audience. Every participant in the show could speak over a live audio channel. The teams were separated into women (aliens) versus men (robots) so that viewers would be able to more easily associate the voices that they heard with the avatars that they saw on the screen. The team members were given cartoon like avatars that could be distinguished by a visible number on their backs and fronts. A speech

bubble would appear above their heads whenever they were transmitting audio. The inhabitants used standard PCs with joysticks and combined headphone/microphone sets. They were located behind the scenes, out of sight of the viewers in the theatre.

These experiments raised a number of significant issues concerning the successful creation of Inhabited TV. One of the most fundamental issues for us is the potential size of Inhabited TV (hundreds of thousands of simultaneous participants) which challenges the scaleability of CVEs and in particular scaleable integration of different media streams. One of our core aims has been to understand and analyse problems of service provisioning so that we can ensure that the necessary resources are available to support a given application or alternatively constrain the application to operate within the limitations of available resources. More specifically, we are interested in identifying the key types of network traffic and determining the bandwidth requirements of Inhabited TV applications used by varying numbers of participants. In addition to this, we also consider the impact of different layers of participation on the network traffic and bandwidth requirements.

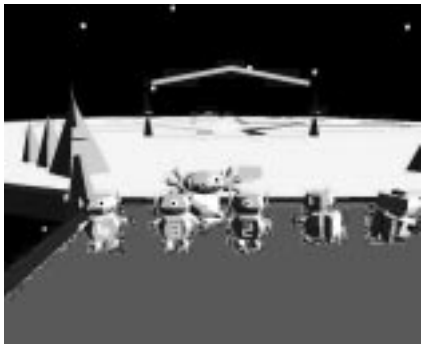

*Fig. 4. The Alien team*

*Fig. 5. Robot captain in falling fish*

## 2. The Method and Data Analysis

Our method involves capturing event logs and traffic measurement data from user trials. The data is then analyzed in three stages. First, statistical analysis of event logs result in a model of user behavior which identifies the average frequencies of occurrences of different events (e.g. moving, speaking) as well as their degree of correlation. Second, we establish a model of application behavior which identifies messages that are generated by a given sequence of events and also the way in which they become concentrated on

different links of a given underlying network topology. Third, we combine the these two models to produce a model of network traffic which predicts peak and mean volumes of traffic on a given link for a given number of participants.

Our data revealed a number of interesting results. For this paper, we concentrate on the results concerning audio which, we believe, has been a neglected aspect of CVEs that have been used for Inhabited TV experiments until now. Our data proved that audio is the most demanding medium: the overwhelming majority of *OOTW* network traffic was multicast of which 91% consisted of peer-to-peer audio! Yet, ways to deal with it in our previous Inhabited TV experiments have either been to completely substitute audio with text (*Heaven & Hell- Live*) or to work with low quality peer-to-peer audio (*OOTW*). In *OOTW*, each participant received an audio mix from other participants' audio streams according to their positions in the world using a combination of attenuation with distance and stereo panning. There has also been little production control over the mixing of audio form the virtual world into the broadcast mix.

Such an approach to audio was justified by the results from our previous analysis of network traffic generated by CVEs. Here, there was found to be very little correlation between different speakers and in some cases even cases of negative correlation (e.g. people tend to avoid speaking at the same time). Contrary to this, our most recent analysis revealed that it could not be assumed that there was at most one speaker at a time for a session. In the *OOTW* network analysis, there was a dramatic increase in a number of participants speaking at the same time (two minutes during which all participants were considered to be simultaneously audio active compared with an uncorrelated expectation of 0.5 seconds). This resulted in significant bursts (peaks) in the audio traffic.

Based on these results, we focused on achieving greater scalability in the audio medium. We started off by investigating the trade-offs between different network topologies for constructing network audio in CVEs suited for future Inhabited TV applications. Our analysis has shown that simple peer-to-peer audio (especially unicast but also multicast), even with silence suppression, becomes increasingly demanding where rates of participation and correlation of audio activity are high. Some techniques for increasing scalability of the audio communication have already been suggested such as: adopting a relatively rigid form of floor control or using audio abstraction techniques (e.g. crowd aggregations provided in our own system MASSIVE-2 mix the audio of their members and re-broadcast a composite audio stream to the rest of the virtual environment [4]).

In this paper we propose and describe two novel approaches to dramatically decrease the network bandwidth over wide area networks in the audio medium. First, we introduce distributed audio servers distributed throughout the network which would aggregate and mix multiple streams in the network. Second, we suggest audio synthesis for generating abstractions for crowds and neighboring virtual environments. We believe that combination of distributed mixing elements and abstraction services will considerably reduce peak system requirements, and thus provide more graceful response to variable patterns of load than current systems would do while retaining open participation (e.g. avoiding floor control).

## 3. Proposed Architecture: Overview

Besides meeting general requirements for audio scalability such as low bandwidth and minimal management, our distributed audio architecture has been particularly designed to support wide area user distribution. We achieve this by preventing an explosion of audio streams over wide area networks. The fundamental idea for minimizing the number of audio streams over WANs is to exploit physical proximity as well as virtual proximity of the users. Our current implementation takes the form of a star network topology of audio servers with the central server at the hub and regional servers running at the outlying nodes (given in Fig. 6). Regional servers are separate and independent audio servers which manage geographically co-located participants e.g. each network site (e.g. LAN) has a regional server to which clients connect statically. The main server has a separate and independent mixer for each remote region, and sends to each region only the mix of streams coming from other region.The functionality of the main server (hub) is given in the Fig. 7. In this way we cancel echo effect over wide area networks. The main benefits for introducing regional servers are:

- Regional servers significantly off-load processing and networking burden from both client and main server computers.
- On average, the total number of audio streams over WAN for this architecture is proportional to the number of different geographical sites. (For the purpose of comparison, the total number of audio streams for multicast architecture would be proportional to the number of simultaneous speakers.) The number of audio streams to and from each regional site is kept constant, since there is only one outgoing (incoming ) audio stream from (to) every site to (from) the hub. This method gives the best results in case when we have a lot of participants coming from relatively few sites.

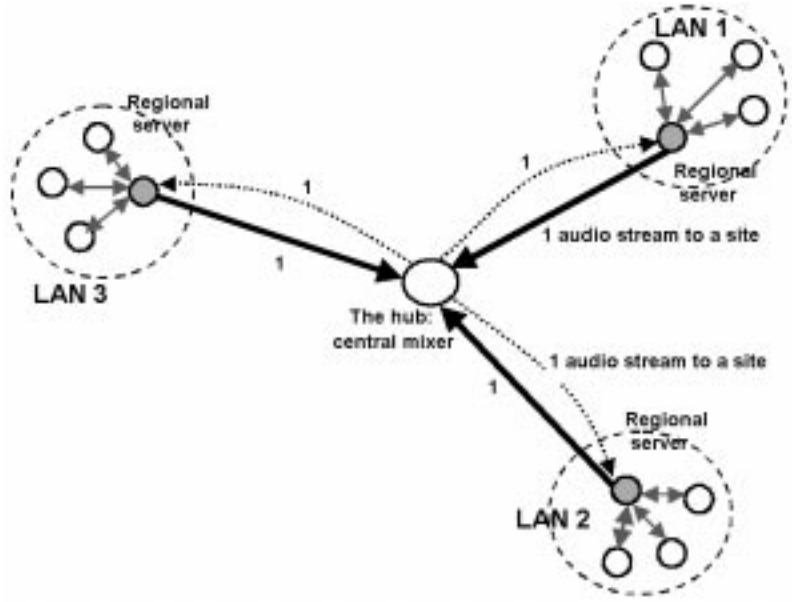


*Fig. 6. Star network topology of audio server*

- In the worst case, when all audio sources are located at different sites, mixing is actually done in the centralized manner.
- Our implementation is designed to minimize latency for geographically co-located speakers (each regional server sends the regional mix straight back to its regional members as well as to the hub) so that RTT for audio packets which originate from geographically close participants is smaller than it would be in the centralized approach (but still larger than for pure peer-to-peer architectures).
- Static connection between clients and regional servers are very good for dial-up connections. This is very important as we expect that in future users will participate in the Inhabited TV shows from their homes.
- Our architecture is very responsive to rapidly changing requirements since it can cope with any number of simultaneous speakers.
- Each distributed component deals only with a subset of users so that the local peaks are correspondingly reduced. The mixing performed in the local audio server limits the impact of the peaks on the rest of the system. In this way both variability in the network traffic and other processing requirements are at least localised if not removed.

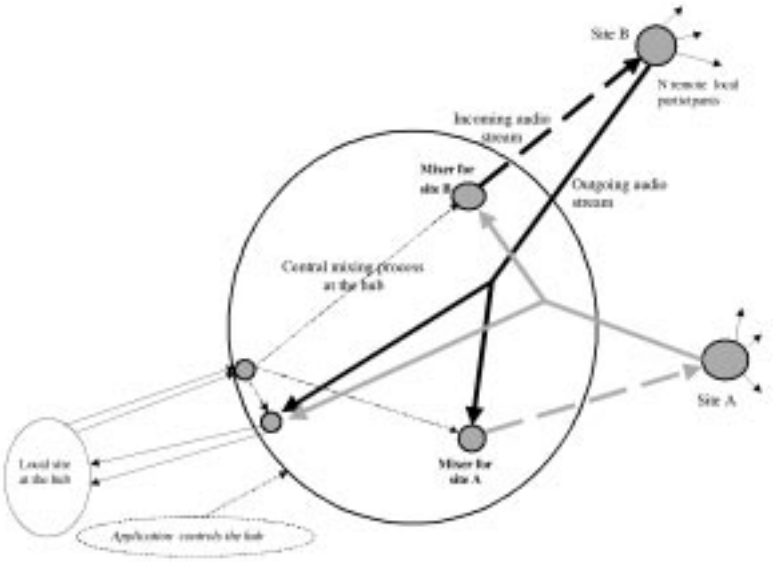


*Fig. 7. Hub Server*

Note that like any shared mixing approaches individual audio spatialisation is not possible. In addition to the distributed star audio architecture, our implementation allows other distributed architecture specifically purely decentralised (peer-to-peer unicast or multicast) and purely centralised architectures (server unicast or multicast)). More specifically, different regions of the virtual world can be configured to have different audio architectures. Since peer-to-peer architectures retain an option for spatialisation, we can have spatialisation in any region which has been configured to have one of the peer-to-peer architectures.

## 4. Discussion and Future Work

### 4.1 Introducing multiple locales

In the analysis till now we have considered what happens when there are many users in one locale only. However since we want to grow the **total population** in the world, we must take multiple locales into consideration and how that impacts the bandwidth. We consider trade-offs in terms of bandwidth between having multiple locales and a single locale for each architecture.

If the players are allocated to several different locales, the total bandwidth over WAN for the peer-to-peer architectures is decreased. However, increasing the number of locales would decrease the performance of the regional distributed architecture. For example, the number of outgoing streams from each site will be equal to the number of different locales people from that site joined. In the worst case when in each site there is a small number of people joined to several locales (for examples $n$), the number of audio streams would be $n \times m \times 2$, where $m$ is number of sites. This amount of streams might saturate low-pass WAN links as well as cause much quicker saturation of the central server CPU load. However, this could be significantly improved by regrouping (reconnecting) the players from several nearby sites: players who were originally connected to the nearest mixing service could be reconnected to some other regional server). In this way we would increase the latency and the bandwidth between nearby regions, but reduce further the bandwidth over WAN. If however there would be many players in each of the different locales connected to the same regional server, then it would not be cost-effective to reconnected all of them to other regional servers. Rather it would be better to make dynamical hierarchical trees of regional servers so that some audio streams get mixed first in the local regional server, then in the nearby regional server before they are sent to the hub. This would further increase latency as mentioned earlier in section 3.2.

### 4.2 Introducing network heterogeneity

Supporting network heterogeneity by combining different architectures: Till now the regional architecture aimed to decrease the total bandwidth in the WAN treating equally all the users, and it was not optimised in any way. At present all the users who joined a certain locale are configured to use one and the same type audio architecture depending on the locale they are in. This means that all users use one and the same shared distribution tree for their audio streams. Our model as well as the implementation is done in such a way to allow easy reuse and extensions. We would like to introduce the capability of providing different routes for some users which would not use the shared distribution tree. Such routes would be optimised according to some metrics for their users and they would be established dynamically. In this way we would support heterogeneity between the users by allowing them to use different audio architectures depending on the network they come from as well as users abstract requirements. For example ATM users who are talking with each other and whom other users must hear with good quality might use peer-to-peer unicast architectures, have direct connections between one another and avoid using the regional server distributed tree.

Dial-up users on the other hand, might use distributed regional architecture. This could be implemented by allowing the users and regional servers to bring in with them the information about their networks and connections. Users and regional servers could be reconfigured either in a centralised or decentralised manner. In case of the centralised way, the application would configure them up appropriately. Alternatively, all the users and regional servers might set themselves up alone based on their abstract requirements and infrastructure constraints.

**Supporting heterogeneity in regional distributed architecture:** This approach could alleviate the problem of no-spatialisation in the regional distributed architecture as the application could instruct the regional servers to treat different users connected to it differently: do the mixing for some users but not for the others: mixing could for example be done only for dial-up users but not for ATM users. Three mixing policies could be optimised: mixing of audio streams only in the local area, outgoing and incoming streams from and to a site. User connections only determine mixing in side the site. The connection between the regional servers and hub would influence whether the mixing of outgoing/incoming audio streams from/to certain region should be performed. For example we might require mixing of incoming streams only and allow no-mixing of outgoing streams so that users with good connections might hear spatialised users with modem connections. If this link is fast, we would need to perform the mixing of outgoing low bandwidth streams so that we allow that other users in other regions.

### 4.3 Sonification

In addition to supporting different architectures for real-time voice communication, our audio service allows on-line sonification of neighbouring environments and crowds. This method involves synthesising an audio stream which represents crowds or environments based on positions, orientations and various activities levels of the members those crowds or environments. This approach minimises the number of audio streams in the network. At present our implementation supports centralised sonification in which the sonified audio is distributed to all interested clients using the audio architecture set up for that specific region of the virtual world. In future, we expect to be able to perform sonification locally on each LAN or host. This will allow us to completely replace WAN audio streams with much less demanding input parameters to the sonification (positions and orientations).

Sonification offers another level of quality of service in the audio medium. It is a complementary approach to our flexible and scaleable framework of architectures and not its alternative. We suggest that sonification

should be most effective when used for the purpose of: awareness (e.g. when we want to tell the people about the things which they can not see) and peripheral awareness (e.g. allow people in one room have some knowledge of what is going on in neighbouring room: how many people there are, how active they are, how co-ordinated their actions are, etc), security and confidentiality (e.g. when we deliberately prevent people from hearing a conversation but still allow them to gain some notion about how many people participate in that conversation and how often much they are moving, etc).

## 5. Conclusion

Inhabited TV aims to create a new entertainment and communication medium by combining traditional TV with CVEs so that the public can become on-line participants within TV shows. Even though there has been a lot of work in designing various techniques and architectures to construct scaleable CVEs, there has been almost no work which focuses specifically on achieving scaleability in the audio medium. However, contrary to previous assumption of negatively correlated audio generated by CVEs used for small group meetings, the results of the *OOTW* analysis have shown that audio can be positively correlated and can dominate network bandwidth requirements in CVE systems used for Inhabited TV applications. We believe that traditional audio architectures used for these applications are neither sufficiently robust to support extremely high-levels of participation nor sufficiently dynamic to cope with very rapidly varying requirements (instant transitions from only few to few hundreds simultaneous speakers). We suggest that scaleable, efficient and flexible audio service is essential for the success of future Inhabited TV applications. This paper has described our audio service which supports traditional centralised and decentralised audio architectures, plus a star multi server audio architecture which performs distributed audio mixing based on the proximity of the users in the physical network topology, and on-line sonification of crowds and environments. This service has been implemented in our new CVE platform which will be used for our next Inhabited TV experiment. We expect that this audio service will allow for truly large scale future Inhabited TV.

## Acknowledgments

Our research was in part funded by British Telecommunications as part of Network Architectures for Inhabited TV. The research concerning sonification was carried out as part of the eRENA project sponsored by European Commision. We would like to thank to Sten Olof for his help and cooperation.

## REFERENCES


1. S. D. Benford, C. M. Greenhalgh, and D. Lloyd: *Crowded collaborative virtual environments.* Proc. CHI'97, Atlanta, USA, 1997, ACM Press, pp. 59-66.
2. C. M. Greenhalgh, S. D. Benford, I. Taylor, M. Bowers, J. M. Walker, G. and Wyver, J.: *Creating a live broadcast from a virtual environment.* accepted as a full technical paper for SIGGRAPH'99, Los Angeles, August, 1999, ACM Press.
3. C. M. Greenhalgh, S. D. Benford, A. Bullock, N. Kuijpers, K. Donkers: *Predicting network traffic for collaborative virtual environments.* Computer Networks and ISDN Systems 30 (1998), 1677-1685.
4. C. M. Greenhalgh, C. And S. Benford: *MASSIVE: A virtual reality system for tele-conferencing.* ACM TOCHI, ACM Press, 1995, pp. 139-261.
5. Radenkovic M., Greenhalgh, C. M., Benford S. D.: *A scaleable audio service for collaborative virtual environments.* UKVRSIG'99, Salford, September, 1999, Springer Press.
6. Greenhalgh, C, Benford S., Taylor I, Bowers J, Walker G, Wyver J,: *Creating a live broadcast from a virtual environment.* to be presented at SIGGRAPH 99, Los Angeles, 9-13 August, 1999.
7. Greenhalgh C, Benford S., Craven M.: *Patterns of network and user activity in an inhabited television event.* to be presented at VRST 99, Los Angeles, 9-13 August, 1999